日本語言い換え処理を利用した日本語–ウイグル語対訳辞書の拡充

小川　泰弘[†]　釜谷　聡史[††]
ムフタル・マフスット[†††]　稲垣　康善[††††]

機械翻訳に対する要求の高まりに伴い，日本語や英語，韓国語といった言語の翻訳に関する研究が進み，実用的なシステムが構築されつつある．その一方で，そうした研究があまり進んでいない言語が存在する．こうした言語においては，翻訳の要である対訳辞書の整備も遅れている場合が多い．一般に対訳辞書の構築には高いコストが必要であり，機械翻訳システムを実現する上での障害となっている．しかし，人間が翻訳作業をする場合，対訳辞書に記載がない単語を別の表現に言い換えて辞書を引くことにより，この問題に対処する場合がある．本研究ではこの手法を模倣し，未登録語を登録語に言い換えることにより対訳辞書を拡充することを提案する．本論文では，対訳辞書の拡充に必要な単語の言い換え処理を収集段階と選抜段階の二つに分割し，前者において語義文に基づく手法を，後者において類似度に基づく手法をそれぞれ適用した．また，類似度に基づく手法では，シソーラスにおける概念間の距離に加え，単語を構成する漢字の語義を利用した．これによって，語法や概念が近く意味的にも等価な言い換えを獲得できた．さらに，獲得した言い換えを翻訳システムで翻訳して日本語–ウイグル語対訳辞書への追加を試みたところ，未登録語 300 語のうち，その 68.3％に対して利用可能な対訳が得られた．

キーワード:　言い換え，機械翻訳，辞書拡張，ウイグル語

Expansion of a Japanese-Uighur Bilingual Dictionary by Paraphrasing

Yasuhiro Ogawa[†] , Satoshi Kamatani[††] , Muhtar Mahsut[†††]
and Yasuyoshi Inagaki[††††]

In machine translation, the number of words in a bilingual dictionary has an important influence on the translation. However, the development cost of such a dictionary is very expensive. In this paper, we resolve this problem by paraphrasing a non-entry word into the entry words. We divide the paraphrasing process into two steps: collecting and screening. In the collecting step, we make paraphrasing expressions of an original word by using its lexical descriptions in a Japanese monolingual dictionary. In the following screening step, we calculate the similarity between the original word and each of its paraphrasing expressions, and choose the best one. We applied this method to our Japanese-Uighur bilingual dictionary. As a result, for 68.3 % of non-entry words, the appropriate Uighur words were given.

KeyWords:　*Paraphrasing, Machine Translation, Dictionary Expansion, Uighur*

1　はじめに

近年，機械翻訳に関する研究が進み，日本語や英語をはじめとし，韓国語，中国語，フランス語など，主要な言語に関してはある程度実用的なシステムが構築されつつある．その反面，そうした研究の進んでいない言語や，機械翻訳の対象となっていない言語が残されているのも事実である．こうした言語においては，言語現象を学習するためのモノリンガル・コーパスや，翻訳知識を得るためのバイリンガル・コーパスなどが充分に蓄積されておらず，また，翻訳の要である対訳辞書の整備も進んでいないことが多い．

そうした，比較的マイナーな言語に関する機械翻訳として，日本語–ウイグル語機械翻訳システム (小川，ムフタル，杉野，外山，稲垣 2000) が研究されている．このシステムにおいては，その原型となった日本語形態素解析システム (小川，ムフタル，外山，稲垣 1999) の日本語辞書が，語彙として約 25 万語，形態素として約 35 万語を収録しているのに対して，日本語–ウイグル語対訳辞書 (ムフタル，小川，杉野，稲垣 2003) は語彙数約 2 万語，形態素数約 3.6 万語[1]と少ないため，翻訳可能な文の数が限られてしまうという問題がある．このように，対訳辞書の規模は，そのシステムが処理できる文数と直接関わる重大な要素である．しかしながら，一般に辞書の構築はコストが高く，登録単語数を増やすことは容易ではない．

これに対して，人間が翻訳作業をする場合を考えると，翻訳者は知らない単語を対訳辞書で検索するが，その単語が辞書に記載されていない場合，同じ意味の別の表現に言い換えて辞書を引く．本研究では，人間のこの行動を模倣し，対訳辞書に登録されていない自立語を，登録されている単語だけから成る表現に言い換えることにより，訳語の獲得を目指す．これにより，二言語間の言語知識が必要な問題を一言語内で扱える問題にすることができる．

言い換えに関する研究は，近年，盛んに進められている (山本 2001)．これに伴って，言い換えの目的に応じた種々の言い換え獲得手法が提案されている．これらの内，本研究で扱う自立語の言い換えに関するものに注目すると，概ね次の二つの手法に分けることができる．一つは，単語の用法や出現傾向，概念などの類似性を評価し，類似する表現を集める手法である (Hindle 1990)(崔，小松，安原 1993) (笠原，松澤，石川 1997)．これらの中には言い換えを獲得することを直接の目的としないものもあるが，集められた類似表現を言い換え可能な語の集合と見做すことができる．もう一つは，国語辞書などにおいて単語の語義を説明している語義文を，その見出し語の意味を保存した言い換えと見做して利用する手法である．これに属する手法としては，語義文から見出し語との同等句を抜き出し，直接言い換える手法 (鍛治，河原，黒橋，佐藤 2002)(釜谷，小川，稲垣 2002) や，2 つの単語間の意味の差を，単語の語義文における記述

† 名古屋大学大学院情報科学研究科, Graduate School of Information Science, Nagoya University
†† 東芝研究開発センター, Corporate Research & Development Center, Toshiba Corporation
なお，この研究は名古屋大学大学院工学研究科在学中に行ったものである．
††† 名古屋大学大学院国際開発研究科, Graduate School of International Development, Nagoya University
†††† 愛知県立大学情報科学部, Faculty of Information Science and Technology, Aichi Prefectural University
1 漢字表記の語彙に対しては，その読みが別の形態素として登録されるため，語彙数と形態素数に差が生じる．

の差異として捉え，言い換えの可否を判定する手法 (藤田，乾，乾 2000)(藤田，乾 2001) が挙げられる．

従来，自立語の言い換え処理は，この二つの分類のどちらか一方の手法を適用して言い換えを得る，一段階の処理として扱われてきた．これに対して，Murata ら (Murata and Isahara 2001) は，言い換え処理を次の二つのモジュールに分割した．一つは，用意した規則を元に，入力表現を可能な限り変換するモジュールであり，もう一つは，変換された表現の内，言い換えの目的に最も適ったものを選び出す評価モジュールである．ただし，変換のための規則は，言い換えの前後で意味が変わらないものであることを保証する必要がある．処理を分割することによって，評価モジュールにおける評価の観点を変えることが可能となり，様々な言い換え目的に対して，汎用的な言い換え処理モデルを提供できるとしている．しかし，この手法では，あらかじめ変換規則を検証しておく必要があるほか，従来の言い換え獲得処理に関する手法を柔軟に適用できないという問題がある．

そこで，本研究では，この言い換え処理の段階分けの考え方をさらに進めて，可能な限り類似表現を収集する**収集段階**と，収集された言い換え候補について，言い換えの目的に適う表現を選び出す**選抜段階**とに分けることを考える．このように分割することにより，各段階において，類似度に基づく手法と語義文に基づく手法とを別々に適用できる．さらに，言い換えの対象となる単語に合わせて，その組み合わせ方を変えることができる．

本論文では，収集段階に語義文に基づく手法を，選抜段階に類似度に基づく手法を用い，両者を組み合わせることによって適切な言い換えを獲得する手法について提案する．さらに，獲得した言い換えを日本語–ウイグル語翻訳システムで翻訳し，それを辞書に追加することによる対訳辞書の拡充実験も行った．

以下，本論文では，第 2 章において現在までに研究されている言い換え処理技術について，その概要を述べて整理する．次に第 3 章において，言い換え処理を収集段階と選抜段階に分割し，それぞれに第 2 章で述べた従来の研究を適用する手法について提案する．第 4 章においては，第 3 章で提案した言い換え手法を用いた実験と，さらに対訳辞書の拡充実験について報告する．最後に，第 5 章は本論文のまとめである．

2　言い換え処理技術の分類

一般に，言い換え処理は「(同一言語内での) 同義表現への言い換え」と捉えることができる．しかし，工学的な言い換え処理を考える場合に，「明示的な意味が同一である表現への言い換え」として捉えると，対象が限定され過ぎてしまう．これに対して，山本 (山本 2001) は言い換え処理を「何かが同一なものへの変換」ではなく「何かの目的を満たす表現への変換」と捉えた．そして，入出力の同一性ではなく，入力表現に対する基準達成の是非に着目し，言い換え処理を「言語表現と換言因子を入力とし，換言因子に沿うように入力表現を変換する処理」と

表 1　換言因子

換言因子	説明
入力誤り訂正	誤りのない表現に
推敲／校正	より自然な表現に
計算機処理に対する頑健化	構文解析に可能な表現に
要約	より短く
詳細化	(計算機／人間にとって) より曖昧さの少ない表現に
簡潔化	易しく分かりやすく
文体	話し言葉／書き言葉に
性別	男言葉／女言葉に
年齢	子ども／高齢者の言葉に
方言	方言に／共通語に
換言因子なし	狭義の換言処理

定義している．ここで換言因子とは，言い換え処理を施す目的であり，山本 (山本 2001) では，表 1 のような例が挙げられている．こうした換言因子ごとに，さまざまな言い換え処理が研究されているが，自立語を類似する別の表現に変換するという点に着目すれば，その手法は，語義文ベースと類似度ベースの二つに大別することができる．各手法について，以下にまとめる．

2.1　語義文ベースの手法

国語辞書の語義文は，見出し語の意味を説明したものであると同時に，意味を充分保存した言い換えであると見ることができる．このような見地に基づく語義文を利用した言い換え獲得手法を，本研究では語義文ベースの手法と呼ぶ．これに属する研究として，次のような例が挙げられる．

- 語義文への直接言い換え

  一般に，国語辞書の語義文には見出し語の意味に加えて，用法・用例なども合わせて記載されている．そのため，そうした余分な記述を削除した言い換えを獲得する必要がある．このための手法として，鍛治ら (鍛治他 2002) は，コーパスを用いて言い換え対象の文と語義文間の格フレームを対応付け，言い換えの際に不必要な格を選定することにより，同等句のみを抜き出してる．

  また，釜谷ら (釜谷他 2002) は，辞書に固有の語義文のパターンに注目し，不必要と考えられる部分を削除するルールを人手により作成し，同等句を切り出している．
- 語義文を利用した意味の差分評価

  藤田ら (藤田，乾 2001) は，単語の語義文の表現の重なりに着目し，単語間の意味の差を評価した．さらに，その重なりの程度を制約とすることで，ある程度良質な言い換えを生成できることを確認している．

2.2 類似度ベースの手法

対象となる単語と他の単語との，共起傾向の類似性や概念的な近さを評価することによって類似度を算出し，それに基づいて類似する単語を集める手法を，本研究では類似度ベースの手法と呼ぶ．こうした研究の多くは，言い換えの獲得よりも類似する単語を集めることを目的としているが，類似する単語グループを言い換え可能な対象と考えることができる．こうした研究には，以下のようなものがある．

- 単語の共起傾向に基づく類似性
  Hindle(Hindle 1990) は，直接評価することが難しい単語間の類似度を，コーパスにおける単語の共起パターンを利用して評価した．この手法は，類似した名詞同士は同じ共起パターンを示すという仮定に基づいている．Hindle は，英語コーパスから共起関係にある主語–動詞–目的語の組を抽出し，その共起傾向の類似性から名詞間の類似度を評価する手法を提案している．
- シソーラス上の概念間距離に基づく類似性
  崔 (崔他 1993) らは，単語の振る舞いや使われ方から見た類似度の尺度として，EDR 日本語概念体系上での概念間の距離から計算した類似度を用いた．その際，注目している単語そのものに付けられている概念だけではなく，その上位概念間の一致も含めた類似度を考慮して類似度を補正している．この手法は，言語に依存しない概念体系を用いるため，任意の言語の任意の二つの単語に対して類似度が計算可能である．
- 概念の知識ベース（概念ベース）に基づく類似性
  笠原ら (笠原他 1997) は，ある単語の国語辞書の語義文中に現れる単語を属性とみることで，その単語の概念を特徴づけた．例えば,「馬」に対する語義文が,「家畜の一．たてがみが長い．草食の動物で…動物….」と与えられた場合,「馬」の概念は，その語義文を構成する「家畜」,「たてがみ」,「動物」などの単語によって特徴付けることができるとした．そして，各単語の概念を，属性を軸としたベクトルによって表現し，二つの単語の概念のベクトルがなす角の余弦を基に類似性を評価している．

3 対訳辞書拡充のための言い換え処理

3.1 本研究における換言因子

本研究の目的は，日本語–ウイグル語対訳辞書の拡充のための言い換えであり，換言因子は「既存の日本語–ウイグル語対訳辞書で翻訳可能，かつ，意味的な過不足の少ない語句への変換」となる．ここで,「既存の日本語–ウイグル語対訳辞書で翻訳可能」という条件は，本研究の最終的な目的である，対訳辞書の拡充に根ざした因子である.「意味的な過不足の少ない」という条件は，本研究によって対訳辞書を拡充した結果，翻訳前後で大きく意味が変わってしまったり，

原文で伝えたい内容が失われることを防ぐための因子である．例えば，「単語数を 漸増させる」を「単語数を 増やす」と言い換えて翻訳した場合，原文における「漸増」の「だんだん」に相当する意味が失われてしまい，本来の内容とは異なる印象を与えてしまう．このような言い換えは，文脈によっては許容することができない場合がある．また，「単語 数を漸増させる」を「言語の最小単位 数を漸増させる」と言い換えて翻訳した場合，言い換えによる意味情報の欠落はないが，回りくどい表現を含んだ違和感のある文章になってしまう．

本研究の目的を満たすためには，意味的な欠落が少なくて，翻訳処理を加えても違和感のない表現を，言い換えとして獲得する必要がある．

3.2　提案する言い換えの枠組み

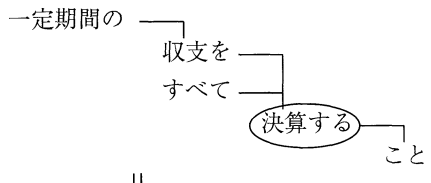
従来の言い換え処理は，前章で述べたように，種々の言語知識からある言い換え表現を獲得する一段階の処理として考えられてきた．しかし本研究では，言い換え処理を，言い換えの候補を集める**収集段階**とその中から不充分あるいは不適切なものを削除する**選抜段階**の二段階に分けて処理する．このように分割することにより，収集段階では言い換え表現の多様性を重視した再現率の高い収集をし，選抜段階では言い換えとして不適格なものを削除する精度を重視した篩い分けをする，といったことが可能となり，多様性と品質に関してバランスの取れた言い換えを獲得することが期待できる．

類似度ベースの手法と語義文ベースの手法は相補的な関係にあることから，本研究では，収集段階と選抜段階の各段階に，類似度ベースおよび語義文ベースの手法をそれぞれ組み合わせて適用することを提案する．類似度ベースの手法は，単語が持つ概念や語の振る舞いの類似性に基づいて言い換えを獲得する手法であり，語義文ベースの手法は，語義文という見出し語と意味的に等価なものを利用して言い換えを獲得する手法である．よって，この二つの手法を組み合わせることで，質の高い言い換えを獲得することが期待できる．そうした手法の内，本論文では，収集段階に語義文ベースの手法を，選抜段階に類似度ベースを用いた手法について述べる．以下，言い換え処理を施す対象となる単語を**言い換え元**，言い換え処理によって得られた語句を**言い換え先**と呼ぶ．

3.3　語義文ベースでの言い換え先候補の収集

本手法では，辞書の語義文を利用して言い換え先の候補を作成する．その際に，2.1節で述べた釜谷ら (釜谷他 2002) の手法を用いる．具体的には，以下の手順で言い換え先候補を収集する．

(1)　語義文を，JUMAN(黒橋，長尾 1999) で形態素解析し，さらにKNP(黒橋，長尾 1998) によって係り受け解析する．これにより，文節情報と係り受け情報を得る．



1. 決算する
2. すべて決算する
3. 収支を決算する
4. 収支をすべて決算する
5. 一定期間の収支を決算する
6. 一定期間の収支をすべて決算する

図 1　「総決算する」の語義の解析結果と獲得される言い換え先候補

(2)　言い換え元が用言の場合は，言い換え先の終わりも用言になるのが望ましいため，語義文末の「こと」「さま」を削除する．

(3)　語義文を構成する文節の組み合わせの内，元の語義文中の係り受け関係を崩さないものを言い換え先候補とする．ただし，語義文の末尾の文節[2]は見出し語を説明する上で主要な役割を果たしているという傾向に基づき，言い換え先候補に必ず含める．

この言い換え先候補収集の手順を，言い換え元「総決算する」とその語義文「一定期間の収支をすべて決算すること」を例として図 1 に示す．まず，手順 (1) において，KNP で係り受け解析がなされ，図 1 の上に示される文節情報および係り受け情報が得られる．さらに，言い換え元「総決算する」は用言であることから，手順 (2) において，文末の “こと” が削除される．結果，言い換え先候補に含まれる可能性のある文節は，「一定期間の」，「収支を」，「すべて」，「決算する」となる．手順 (3) で，「決算する」を含む全ての文節の組み合わせが作られる．ただし，元の係り受け関係が崩れてしまうような組み合わせである，「一定期間の決算する」などは候補から外す．以上により図 1 中の下に示す，6 個の言い換え先候補が獲得される．

3.4　語の構成に基づく意味推定と意味因子

前節で収集した言い換え先候補から，言い換えとして意味的に過不足のない言い換え元の同等句を選び出す．そのためには，どの候補が最も適切な言い換えであるか，その指標を定めなければならない．

例として，言い換え元「点火する」に対して，「灯す」「火を灯す」「物に火を灯す」という三つの言い換え先候補がある場合を考える．これらの中で言い換え元の同等句として相応しいのは，「火を灯す」である．日本語の単語の意味は，そこに含まれる部品の意味から構成されている場合が多い．「点火する」の例では，漢字 “点” の意味「点ける」と，漢字 “火” の意味「火」

2 ただし，「こと」「さま」は，(2) で除去されているので，それを除いた末尾となる．

から「火を灯す」という語義が構成されているといえる．このように，その単語の意味を構成している意味の部品とも言うべきものを，本研究では**意味因子**と呼ぶ．

本研究では，[**漢字**]，[**部分**]，[**全体**] の 3 種類の意味因子を定義し，言い換え元となる単語の意味は，この意味因子のいくつかの組み合わせで表現されていると考える．各意味因子の具体的な定義は以下の通りである．

漢字　言い換え元を構成している漢字一字ごとの意味

漢字は表音文字であると同時に表意文字であるから，漢字によって構成された単語は，その漢字と関係のある語義を内包していると考えられる．先に例に挙げた「点火する」の意味は，漢字 “点” の意味と漢字 “火” の意味の合成によるもの考えられ，これらを最小の意味因子としてみることができる．

部分　言い換え元の意味のある部分

例えば,「一括払い」という語は,「一括」と「払い」の二つの単語からその意味が構成されており，そうした言い換え元の一部が意味因子となる場合もある．

全体　言い換え元そのもの

言い換え元全体で，その意味を表す場合がある．例えば,「右往左往する」という単語は，単純な漢字の語義の組み合わせで語義が構成されているのではなく，組み合わせたことによって，新たに「混乱する」といった意味が生まれたと考えられ，これ全体が意味因子である．

実際には，言い換え元がどの意味因子から構成されているかを求める必要がある．そのため，あらかじめ意味因子の候補を可能なだけ集め，言い換え先候補と比較することで意味因子を決定する．その際，[部分] と [全体] については，EDR 日本語単語辞書 (日本電子化辞書研究所 1996) を引き，そこに記載されている概念識別子を各意味因子候補の概念識別子とする．多義語の場合には，複数の概念識別子が存在するが，そのすべてを利用する．そして，3.6 節でこの概念識別子を利用して類似度を計算し，言い換え元がどの意味因子から構成されているかを決定する．なお，意味因子 [部分] の場合，例えば「一括払い」における「括払」のように，EDR 日本語単語辞書に掲載されていないものは意味因子の候補とはならない．また，EDR 日本語単語辞書において名詞とサ変動詞 (例えば「決算」と「決算する」) が区別されているため，意味因子候補として両方を考える．さらに，意味因子 [漢字] の概念識別子は，次節で説明する漢字意味辞書に基づいて決定する．

同様に，言い換え先に含まれる各自立語についても EDR 日本語単語辞書を引き，その概念識別子を求めておく．図 2 に，言い換え元「総決算する」に対する意味因子の候補と，その言い換え先候補「一定期間の収支をすべて決算する」に含まれる自立語の概念識別子を示す．なお，これ以降，言い換え元と言い換え先のペア，例えば「総決算する→一定期間の収支をすべて決算する」を**言い換え対**と呼ぶ．この例では，実際に意味因子となるのは,「決算する [部分]」

観点	語	概念識別子
言い換え元		
[全体]	総決算する	0fa9e8, 0faa02
[部分]	総	106cb1, 3bf848, 3cf3a0, 0ea7e0, 0fa8e0
	決	0ef621, 3ce68e, 3ce80c, 3ce93c, 3cf0f2, ...
	決算	3c3b1d, 3c3b1e, 3c3b1f, 0ef51f, 0ef520, ...
	決算する	0ef51f, 0ef520
	算	3cf83a, 3cf83a, 0f37b0, 0f37b1, 3ce988, ...
	算する	3cf060
[漢字]	総, 決, 算	— (漢字意味辞書中の語義に従う)
言い換え先候補		
言い換え先候補中の自立語	一定期間	1f9de1
	収支	3c4225
	すべて	0e472c, 3d04f3
	決算する	3c3b1d, 3c3b1e, 3c3b1f, 0ef51f, 0ef520, ...

図 2　「総決算する → 一定期間の収支をすべて決算する」における概念識別子

と,「総 [漢字]」である．その求め方については，3.6 節以降で説明する．

3.5　漢字意味辞書

前節で述べた意味因子 [漢字] を使用するため，漢字の語義を記述した漢字意味辞書を広辞苑第四版 (CD-ROM 版)(新村 1996) から以下の手順で構築した．

(1)　漢字一文字ごとに，その読みを区別せずに語義文を取り出す．

(2)　意味番号，出典，用例など，辞書特有の付記情報を削除する．

(3)　句点「.」ごとに語義文を分解し，ぞれぞれを一つの語義とする．例えば，語義文に「思慮．おもわく.」とあれば,「思慮」と「おもわく」を語義とする．

(4)　辞書特有の表現である文末の「さま」,「こと」を削除する．

(5)　「もの」,「人」及び，それに係る語を削除する．これは，これらの語が一般的であり，また，説明を補足するために用いられることが多いからである．

(6)　削除した結果，文として不自然になったものを人手によって修正する．その際，語義に関する部分には手を加えず，あくまで不自然な部分の修正に留めた．

(7)　それぞれの語義を，JUMAN と KNP を利用して係り受け解析する．

表 2　漢字意味辞書の一例

漢字見出し	語義	語義品詞	概念識別子
総	総	名詞	106cb1, 3bf848, 3cf3a0
	糸	名詞	0e4dd9, 0e4ddc, 0e4ddd, ...
	束ねる	動詞	0fc4e7, 3ce6ce, 3ce7de, ...
			⋮
決	決める	動詞	0ec4a7, 0ec4a9, 0ef563, ...
	思い切る	動詞	0e80d8, 0ef623
	可否	名詞	0ea373, 0ea374, 0ea375
	定める	動詞	0e7749, 0ec249, 0ef563, ...
			⋮
算	数	名詞	0e998b, 0e998f, 0f87d9, ...
	勘定	サ変名詞	0eaeed, 0eaeed, 0eaeef, 3cee38
	数	名詞	0e998b, 0e998f, 0f87d9, ...
	数える	動詞	0e9ae2, 3cf060
			⋮

(8)　語義を構成する各自立語について，EDR 日本語単語辞書 (日本電子化辞書研究所 1996) に示された概念識別子を意味因子 [漢字] の概念識別子とする．

なお，意味因子 [全体] および [部分] の場合と同様に，多義語には複数の概念識別子が存在するが，それらをすべて利用する．本来ならば，類似度を求める際にその単語がどのような概念で使われているかを定め，それに基づいて計算する必要がある．しかし，こうした概念の特定は煩雑でコストがかかることから，本研究では，そうした特定も次の選抜段階で行うこととし，意味因子の候補を挙げる際には，すべてを列挙した．

以上の作業によって作成した漢字意味辞書の一部を表 2 に示す．ここで，漢字に複数の語義があれば，それぞれについて項目を用意する．また,「総」に対する「糸/束ねる」や「決」に対する「可否/定める」のように，一つの語義が複数の自立語から構成される場合もある．

3.6　意味因子候補と言い換え先候補の類似度

3.4 節で述べた意味因子の候補と，言い換え先候補に含まれる自立語との間の類似度を，単語間の類似度に基づいて計算する．ここで単語間の類似度は，その単語がもつ概念間の距離に基づいて計算する．具体的には，シソーラスの一種である EDR 概念体系辞書 (日本電子化辞書研究所 1996) を用いて，対象となる単語が持つ概念間距離を長尾 (長尾 1996) によって紹介された以下の式 (1) をベースにして計算する．

$$\frac{2 \times \mathrm{depth}(\mathrm{csc}(c_1, c_2))}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)} \tag{1}$$

ここで，$\mathrm{csc}(c_1, c_2)$ は，二つの概念 c_1 と c_2 のシソーラスにおける共通上位概念を，$\mathrm{depth}(c_1)$ は概念 c_1 の根からの深さを示す．なお，根の深さを 0 とするため，この式における最小値，す

表 3　意味因子「糸/束ねる [漢字 “総”]」に対する類似度

漢字 “総” の語義に含まれる自立語	言い換え先の自立語			
	一定期間	収支	すべて	決算する
糸	0	0.29	0.18	0
束ねる	0	0	0	0.43

なわち，まったく関係のない概念間の類似度は 0 になる．

木構造をしているシソーラスの場合，共通上位概念および，そこへの経路が一つに定まるが，EDR 概念体系辞書は多重継承を許し，一つの概念に二つ以上の上位概念が存在するため，共通上位概念とそこへの経路が複数考えられる．その場合には，類似度が最も高くなる共通上位概念と経路を定め，そのときの値を採用する．また EDR 日本語辞書においては，多義語には複数の概念が付与されているが，その場合にも値が最大になる概念を選ぶ．よって，単語 w_1, w_2 が，それぞれ複数の概念 $c_{11}, \cdots, c_{1i}, \cdots$ および $c_{21}, \cdots, c_{2j}, \cdots$ をもつ場合，その類似度 $SIM(w_1, w_2)$ は以下の式 (2) のように計算される．ただし，$\mathrm{csc}_k(c_i, c_j)$ は概念 c_i, c_j の複数ある共通上位概念の一つを示すものとする．

$$SIM(w_1, w_2) = \max_{i,j,k} \frac{2 \times \mathrm{depth}(\mathrm{csc}_k(c_{1i}, c_{2j}))}{\mathrm{depth}(c_{1i}) + \mathrm{depth}(c_{2j})} \tag{2}$$

ここで意味因子 [全体] と [部分] は，1 単語で表現されるから，各意味因子 m_i と言い換え先 p に含まれる自立語 p_j と間の類似度が上記の $SIM(m_i, p_j)$ で計算できる．しかし，意味因子 [漢字] については，前節で述べたように，漢字意味辞書において語義が複数の自立語からなる場合があり，その場合には，類似度を直接計算することができない．そこで，意味因子 m_i が n 個の自立語 $m_{i1}, \cdots, m_{in}$ から構成される場合は，意味因子 m_i と言い換え先候補に含まれる自立語 p_j との類似度を以下のように計算する．

$$SIM'(m_i, p_j) = \begin{cases} 0, & \sum_k SIM(m_{ik}, p_j) = 0 \text{ のとき} \\ \sqrt[n]{\prod_k^n \max_j SIM(m_{ik}, p_j)}, & \text{それ以外} \end{cases} \tag{3}$$

すなわち，意味因子 m_i を構成するすべての単語 m_{ik} に対して $SIM(m_{ik}, p_j) = 0$ となる場合には類似度を 0 とし，そうでない場合には，各 m_{ik} に対して $SIM(m_{ik}, p_j)$ が最大になる場合を求め，その相乗平均を類似度とする．なお，この式においては，すべての m_{ik} について $SIM(m_{ik}, p_j) = 0$ になる場合以外は，どの p_j に対しても，類似度 $SIM'(m_i, p_j)$ は同じ値となる．

例えば，意味因子「糸/束ねる [漢字 “総”]」の場合を考えると，この意味因子は，二つは自立語から構成されている．まず，意味因子 [漢字] を構成する各自立語と，言い換え先の

表 4　「総決算する」における類似度計算

意味因子候補	言い換え先の自立語			
	一定期間	収支	すべて	決算する
総決算する [全体]	0	0	0	0.2
決算する [部分]	0	0	0	1.00
決算 [部分]	0	0.88	0	0
算する [部分]	0	0	0	0.66
⋮	⋮	⋮	⋮	⋮
すべて [漢字 “総”]	0	0	1.00	0
糸/束ねる [漢字 “総”]	0	0.35	0.35	0.35

各自立語との間の類似度 SIM を計算すると，表 3 のようになる．ここで，$SIM(糸, p_j)$ と $SIM(束ねる, p_j)$ の最大値は，それぞれ $SIM(糸, 収支) = 0.29$，$SIM(束ねる, 決算する) = 0.43$ となり，この相乗平均 $\sqrt{0.29 \times 0.43} = 0.35$ が $SIM'(糸/束ねる\ [漢字\ “総”], 収支)$ の値となる．同様に $SIM'(糸/束ねる\ [漢字\ “総”], すべて) = SIM'(糸/束ねる\ [漢字\ “総”], すべて) = 0.35$ となる．ただし，$SIM(糸, 一定期間) = SIM(束ねる, 一定期間) = 0$ のため，$SIM'(糸/束ねる\ [漢字\ “総”], 一定期間)$ の値だけは 0 になる．

以上の方法に基づいて，言い換え元「総決算する」における類似度を計算したものを表 4 に示す．ここで類似度とは，意味因子が 1 単語で構成される場合は SIM，2 単語以上で構成される場合は SIM' の値である．

3.7　言い換え先の選抜

前節で定義した類似度を用いて，言い換え元における意味因子の決定と，言い換え先の選抜を行う．なお，本手法では，前節の類似度 SIM' の計算や，これ以降の計算においても，平均を求める際には相加平均ではなく相乗平均を用いる．これは，類似度を用いる目的が選抜であり，不適切な候補をできるだけ篩い落とすことに主眼を置いているからである．よって，平均を計算する対象中に，値の低いものがあると平均値がより低くなる相乗平均を用いた．

意味因子の決定

言い換え先の各自立語に対して，類似度が最大となるものを，それぞれ意味因子とする．よって表 4 の例では，「収支」に対して「決算 [部分]」(類似度 0.88)，「すべて」に対して「総 [漢字]」(類似度 1.00)，「決算する」に対して「決算する [部分]」(類似度 1.00)，がそれぞれ対応する意味因子となる．なお，「一定期間」は，すべての意味因子候補との類似度が 0 なので，対応する意味因子が存在しないとする．ただし，以下の計算において「一定期間」と意味因子候補との間の類似度が必要になった場合には，その値を 0 と見做す．

言い換えの効率性

言い換え先の選抜基準を考えると，まず言い換え元の語をなるべく少ない単語数で表現し，冗長な表現を含まないものが望ましい．これを**言い換えの効率性**と定義する．本手法の枠組では，言い換え元に含まれる意味因子を表すのに必要のない単語が含まれていないものが良い言い換え先となる．この効率性を計算するために，言い換え元の意味因子との類似度を，言い換え先候補に含まれる各自立語の**有用度**と考え，その相乗平均を言い換え先候補全体の**効率** *Eff* とする．例えば，*Eff*(収支をすべて決算する) $= \sqrt[3]{0.88 \times 1.00 \times 1.00} = 0.96$ となるが，*Eff*(一定期間の収支をすべて決算する) の場合は，有用度 (類似度) が 0 となる自立語「一定期間」を含むため，その値が 0 となる．すなわち，*Eff* は，言い換え先が有用度の高い語だけで構成されているほど 1 に近い値を，有用度の低い語を含むほど 0 に近い値をとる．

言い換えの充足性

効率性とは逆に，言い換え元の意味がすべて言い換え先に含まれているかどうかを判定する，**言い換えの充足性**を考える必要がある．本手法の枠組では，言い換え元の意味は，それを構成している漢字から成ると考えている．よって，漢字一字ごとに，その意味が言い換え先にどの程度反映されているかを示す**反映度**を考える．この反映度は，その漢字を含む意味因子と，対応する言い換え先に含まれる自立語との間の類似度とし，各漢字ごとの反映度の相乗平均を言い換え先の**充足率** *Suf* とする．

例えば，*Suf*(すべて決算する) は，言い換え元「総決算」を構成する各漢字に対し，「総」を含む意味因子「総 [漢字]」の反映度 1.00，「決」を含む意味因子「決算する [部分]」の反映度 1.00，「算」を含む意味因子「決算する [部分]」の反映度 1.00 の相乗平均として求められ，その値は 1.00 となる．一方，*Suf*(決算する) の場合は，「総」を含む意味因子に対応する自立語が言い換え先「決算する」に存在しない．この場合，「総」の反映度を 0 とし，結果，*Suf*(決算する) $= 0$ となる．すなわち，言い換え先に言い換え元の意味を表す自立語が抜けていると充足率 *Suf* の値は 0 となる．逆に，言い換え先に余分な自立語がある場合，例えば，*Suf*(収支をすべて決算する) の値は *Suf*(すべて決算する) と同じく 1.00 となる．

言い換えの妥当性

本手法における，妥当な言い換えとは，言い換えの効率性と言い換えの充足性の両方が満たされているものである．よって，効率 *Eff* と充足率 *Suf* の二つの値の相乗平均を，**言い換えの妥当性** V とする．言い換え元「総決算する」に対する例では，V(収支をすべて決算する) $= \sqrt{0.96 \times 1.00} = 0.98$，$V$(すべて決算する) $= \sqrt{1.00 \times 1.00} = 1.00$ となり，「すべて決算する」が最も妥当な言い換えといえる．

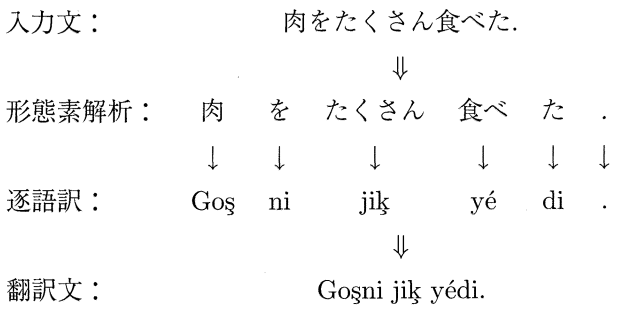


図 3　日本語–ウイグル語機械翻訳

4　日本語–ウイグル語対訳辞書拡充実験

4.1　日本語–ウイグル語機械翻訳

日本語とウイグル語は共に膠着言語であり，語順がほぼ同じであるなど構文的にも類似した点が多い．こうした共通点に着目し，両言語を共に派生文法 (清瀬 1989) で記述し，日本語入力文を形態素解析し，その後逐語訳することでウイグル語訳文を生成する方法 (図 3) が (小川他 2000) において提案されている．

この日本語–ウイグル語機械翻訳システムでは，入力された日本語文に対する形態素解析結果を逐語訳することを基本としているが，形態素解析に用いる辞書の登録単語数が約 35 万語であるのに比べて，対訳辞書に登録されている単語は約 3.6 万語と少ないため，形態素解析は可能であるが翻訳をすることができない事例が多く見られる．

本研究では，第 3 章で提案した言い換え獲得の枠組みを用い，未登録語を言い換えて，それを翻訳することによって対訳語を自動的に獲得し，辞書拡充を図る実験を行った．具体的には，未登録語の中からコーパスにおいて出現頻度の高い単語を第 3 章の方法で言い換えて，その獲得された日本語の言い換えを，(小川他 2000) の日本語–ウイグル語機械翻訳システムを用いて翻訳した．そして，言い換え結果が完全に翻訳できた場合に，言い換え元の訳語として辞書に登録できるかどうかを判定し，言い換え処理が対訳辞書の拡充にどのように寄与するかを検証した．

4.2　言い換え先候補の収集

名詞，動詞，サ変名詞の各品詞ごとに，EDR 日本語コーパス (日本電子化辞書研究所 1996) における出現頻度が上位 1,000 位までとなる単語を収集し，そのうち，日本語–ウイグル語辞書に登録されていなかった名詞 452 個，動詞 477 個，サ変名詞 396 個を実験対象とした．ただし EDR 日本語単語辞書では，各単語について概念が異なれば別エントリとして登録しているた

表 5　言い換え先候補

品詞	単語数	言い換え元数	言い換え先候補数	最大候補数	最小候補数
名詞	452	473	2,897	129	1
動詞	477	514	2,541	82	1
サ変名詞	396	429	2,087	55	1

め，言い換え元として考えるときには概念の異なりごとに区別した．また，複数の読み方がある場合や表記が異なる場合にも区別した．結果，名詞 473 概念，動詞 514 概念，サ変名詞 429 概念を実験対象とする言い換え元とした．

そして，各言い換え元に対する EDR 日本語単語辞書の概念説明を語義文と見做し，3.3 節で述べた手法により言い換え先候補を収集した．結果，表 5 に示す言い換え先候補が収集された．

4.3　言い換え先候補の選抜

前節で収集した言い換えを，3.7 節に示した類似度ベースの手法で選抜する．

その際，3.5 節で述べた漢字意味辞書が必要になる．本来はすべての漢字について辞書を作成するのが望ましいが，人手による修正が必要な部分があるため，今回は実験に必要な漢字についてのみ作成した．その結果，漢字 605 文字に対して，合計 10,502 個の語義をもつ辞書を作成した．漢字 1 字あたりに付加された語義の平均は 17.3 個であり，付加された語義数の最大は 113 個，最小は 1 個であった．

そして，この漢字意味辞書に基づく意味因子 [漢字] と，EDR 日本語単語辞書の概念識別子に基づく意味因子 [部分]，[全体] を 3.7 節に示した手法で選抜した．すなわち，各言い換え元に対して言い換えの妥当性 V の値が最も高いものを言い換え先とした．ただし，妥当性 V の値が最大となるものが複数あった場合，言い換えとして同等に適切であると考え，一つの言い換え元に対して複数の言い換え先があるとした．よって，評価する言い換え対 (言い換え元と言い換え先のペア) も一つの言い換え元に対して複数存在する場合がある．以上の結果，それぞれの品詞ごとに名詞 537 個，動詞 599 個，サ変名詞 477 個の言い換え対を得た．

4.4　評価

得られた言い換え対のうち，各品詞 300 個をランダムに抜き出し，日本語の言い換えとして適切であるかどうかを人手で評価した．その際，以下の観点に基づいて結果を分類した．

まず，言い換え成功としたものを，以下の二つに分類した．

妥当　適切に言い換えられたもの．

文脈依存　言い換えが適切であるかどうかが文脈に依存するもの．例えば，「参画する → 計画に加わる」という言い換えは，「彼も 参画し ている」という文脈では適切であるが，「経

表 6　日本語の言い換えとしての評価

評価基準		名詞		動詞		サ変名詞	
言い換え成功	妥当	80	(26.7%)	144	(48.0%)	146	(48.7%)
	文脈依存	43	(14.3%)	38	(12.7%)	47	(15.7%)
言い換え失敗	説明過剰	23	(7.7%)	21	(7.0%)	22	(7.3%)
	意味欠落	37	(12.3%)	31	(10.3%)	18	(6.0%)
	その他	85	(28.3%)	29	(9.7%)	10	(3.3%)
	国語辞書	32	(10.7%)	37	(12.3%)	57	(19.0%)
計		300	(100%)	300	(100%)	300	(100%)

営に参画している」と文脈では不適切となる．

一方で，言い換え失敗としたものを，以下のように分類した．

説明過剰　言い換え先が言い換え元の説明になっていて，言い換えとしては記述が過剰であるもの．例えば，「本土 → その国の中心をなす国土」という言い換えが，これに分類される．

意味欠落　言い換え元の意味の一部が欠落した言い換え先が得られたもの．例えば，「苦戦する → 戦いをする」という言い換えがこれに当たる．

国語辞書の不充分な記述　本実験ではEDR日本語単語辞書の概念説明を語義文として利用したが，この概念説明は，人間が他の概念と区別するためのものであり，単語の説明となっていないものがある．そうした語義文の記述に由来する失敗はこれに分類される．

その他　上記以外のもの．この中には選抜手法に原因が求められるものが多く，その点については考察で言及する．

こうした基準で評価した結果を表6に示す．

4.5　日本語の言い換えに関する考察

表6において，言い換えに失敗していると判定されたものについて検討する．まず，国語辞書の記述が不充分であったために誤りとされたものが多い．EDR辞書の概念説明は，一般の国語辞書における語義文とは異なり，人間が概念を区別する際の参考とするためのものであり，概念の説明中で見出し語をそのまま用いている場合がある．例えば，サ変名詞「会話する」の語義文は，そのまま「会話する」となっており，実験では，「会話する → 会話する」という言い換えが得られたが，これは言い換えとしては失敗である．こうした点については，他の辞書の語義文を用いることで改善されると考えられる．

次に，各品詞ごとに検討する．まず名詞については，表6に示したように，動詞やサ変名詞に比べて言い換えに成功した割合が低い．これは，名詞に対する国語辞書の語義文は，その見出し語がどのようなものであるかを説明している傾向が強いからである．例えば，「売り場」に対する語義文は，「物を売る一定の場所」となっており，このような語義文からは，言い換えとしての同等句を取り出しにくい．よって，名詞に対しては本手法とは逆に，収集段階で類似度

表 7　意味因子 [漢字] の利用による言い換え結果の変化

		言い換えの品質				
		向上	同じ	同等	低下	計
妥当性 V の値	増加	85	116	36	12	249
	変化なし	5	169	0	6	180
	減少	0	0	0	0	0
計		90	285	36	18	429

ベースの手法を，選抜段階で語義文ベースの手法を適用することによって，より効果的な言い換えが得られると予想される．

一方，用言の言い換えでは，例えば「一体化する」に対して「まとめる」が語義文となっているように，比較的易しい言葉で言い換えられるものが多い．そのことから，今回用いた語義文ベースの収集と，類似度ベースの選抜を組み合わせた手法は，用言向きの手法であるといえる．実際，用言（動詞とサ変名詞）についての結果を見れば，用言の言い換えは，辞書に起因する誤りを除けば，7〜8 割の精度で言い換えに成功しており，本手法の有効性を示す結果であるといえる．

また，その他に分類した例では，選抜段階での失敗が挙げられる．実験では「落ちつく → なる」という言い換えが得られたが，これは「落ちつく」の語義文「心が安定した状態になる」から得られたものである．本手法は，言い換え元の意味因子と対応する自立語を，係り受けを保ったまま語義文から切り出す．この例では，「状態」という語と対応する意味因子がなかったために，「安定した」という文節が抜き出せなかったものと考えられる．

同様の例として，「書ける → ことができる」などの可能の意味含む動詞に関しては，ほぼ全ての事例で誤った結果を得ていた．例えば，「書ける」の語義は「書くことができる」であるが，言い換え先としては「できる」が獲得されていた．これは，本手法が元の係り受け関係を保存して言い換え先候補を収集し，選抜することに起因する．この例では，「こと」という単語の有用度が低く計算されたために，これに係る「書く」も削除されてしまったと考えられる．この問題の解決策としては，頻出単語や抽象度の高い単語に関しては類似度計算の対象としない手法が考えられる．

4.6　意味因子 [漢字] の有効性に関する実験と考察

ここで，本研究の特徴である意味因子 [漢字] の有効性について検討する．そのために，4.4 節の評価で最も結果の良かったサ変名詞の言い換えに対して，意味因子 [漢字] を利用した場合と利用しなかった場合の比較実験を行った．

まず，言い換え元となるサ変名詞 429 概念のそれぞれについて，意味因子 [漢字] を利用しない場合をベースとし，意味因子 [漢字] を利用することにより，妥当性 V の値と得られる言い換え先がどのように変化するかを実験で確かめた．その結果を表 7 に示す．ここで，言い換えの

表 8　サ変名詞の言い換えにおける意味因子 [漢字] の有無による比較

評価基準		[漢字] あり		[漢字] なし	
言い換え成功	妥当	146	(48.7%)	116	(36.4%)
	文脈依存	47	(15.7%)	41	(12.9%)
言い換え失敗	説明過剰	22	(7.3%)	18	(5.6%)
	意味欠落	18	(6.0%)	77	(24.1%)
	その他	10	(3.3%)	13	(4.1%)
	国語辞書	57	(19.0%)	54	(16.9%)
計		300	(100%)	319	(100%)

品質が「同じ」とあるのは，得られた言い換え先が同じであることを示す．一つの概念から複数の言い換え先が得られる場合は，それぞれが一致したものをこれに分類する．得られた言い換え先が異なった場合は，それが意味因子 [漢字] を利用しなかった場合と比べて，向上しているか，同程度であるか，低下しているかで判定した．

妥当性 V の値が増加し，言い換えの品質が向上した 85 個の中には，意味因子 [漢字] なしの場合には，すべての言い換え先候補について妥当性の値が 0 となり言い換え先が得られなかったが，意味因子 [漢字] を用いることで言い換え先が獲得できたものも含まれている．今回の実験では，そうしたものが 16 個あった．

表 7 をみると，意味因子 [漢字] を利用することにより，多くの場合に妥当性 V の値が増加することが分かる．これは，本手法では，言い換え先の各自立語に対して類似度を最大とするものを意味因子とするからである．つまり，意味因子 [漢字] を利用することにより，意味因子候補がその分増えることになる．そうした候補と言い換え先の各自立語との類似度が，他の候補よりも低ければ意味因子として選ばれず，結果は変化しない．しかし，その類似度が高ければ意味因子として選択されることになり，最終的な妥当性 V の値が向上する．なお，以上の理由により，意味因子 [漢字] を利用することによって選択される意味因子が変化しても，妥当性 V の値が減少することはない．これは実験結果からも確かめられた．

さらに，言い換えとしての評価について表 6 に示したサ変名詞の分と，意味因子 [漢字] を利用しなかった場合との結果を合わせて表 8 に示す．今回は妥当性 V の値が最大となったものを評価し，最大となるものが複数ある場合は，そのすべてを評価した．ただし，意味因子 [漢字] の有無で妥当性の値が変化し，[漢字] なしのとき最大となるものが複数あったが，[漢字] ありの場合には一つになったものがある．そのため，言い換え元のサ変名詞は同じものを利用したが，得られた言い換え先の総数が異なっている．

表 8 の結果から，[漢字] ありの場合に比べて，[漢字] なしの場合には，妥当なものが減り，意味欠落と評価されたものが多くなっていることが分かる．これは表 4 を見ると理由が理解しやすい．表 4 の例において，意味因子 [漢字] を用いなかった場合，その他の意味因子候補と言い換え先の自立語「すべて」との類似度がすべて 0 になり，自立語「すべて」に対応する意味因

表 9　翻訳結果

	名詞		動詞		サ変名詞	
翻訳成功	245	(81.7%)	273	(91.0%)	221	(73.7%)
翻訳失敗	48	(16.0%)	16	(5.3%)	68	(22.6%)
解析失敗	7	(2.3%)	11	(3.7%)	11	(3.7%)
合計	300	(100%)	300	(100%)	300	(100%)

表 10　対訳語としての適切さ

品詞	適切		条件付		不適切		合計	
名詞	37		26		37		100	
動詞	41		26		33		100	
サ変名詞	41		34		25		100	
合計	119	(39.6%)	86	(28.7%)	95	(31.7%)	300	(100%)

子が存在しなくなる．そのため，自立語「すべて」を含む言い換え先の効率 *Eff* が下がり，そうした言い換え先が選択されなくなる．そして，結果的に言い換えとして意味が欠落したものが得られることになる．こうした点からも，意味因子 [漢字] を利用することの有用性が分かる．

また，今回は各言い換え元ごとに妥当性の値が最大となるものを選んだが，特定の閾値を用いて選抜する場合には，妥当性 V の値がある程度の大きさをもつ必要があり，そうした場合にも意味因子 [漢字] の利用は重要となる．

4.7　対訳辞書の拡充

4.4節で評価した，各品詞300組の言い換え対について，その言い換え先を日本語–ウイグル語翻訳システム (小川他 2000) によって翻訳した．翻訳の成否に関する結果を，表9に示す．ここで，翻訳成功としたものは，言い換え先が全てウイグル語に変換されたものであり，翻訳失敗としたものは，言い換え先の単語の中に日本語–ウイグル語翻訳システムで使用した辞書に登録されていない単語が含まれていたものである．解析失敗としたものは，日本語–ウイグル語翻訳システム (小川他 2000) が入力文の解析に失敗し，出力を得られなかったものである．

さらに，言い換え先の翻訳結果が，言い換え元の訳語として適切かどうかをウイグル語ネイティブによって評価した．評価対象は，それぞれの品詞につき翻訳が成功したもののうち，各100単語をランダムに取り出したものである．その結果を表10に示す．表中で,「条件付」としたものは，常にその表現を用いることができるわけではないが，文脈によっては認められると判定されたものである．

4.8　辞書拡充に関する考察

表9から，言い換えた結果が，おおむね翻訳可能であることが分かる．このことから，未登

表 11　日本語言い換えとしての適切さと対訳語としての適切さとの対応

		名詞			動詞			サ変名詞		
		適切	条件付	不適切	適切	条件付	不適切	適切	条件付	不適切
言い換え成功	妥当	14	7	6	24	12	9	25	18	14
	文脈依存	5	5	3	4	5	4	6	12	2
言い換え失敗	説明過剰	6	1	2	1	1	8	6	2	0
	意味欠落	2	6	6	9	3	3	2	1	2
	その他	6	6	17	1	1	8	1	0	3
	国語辞書	4	1	3	2	4	1	1	1	4
計		37	26	37	41	26	33	41	34	25

録語を言い換えることの有効性が示せた．また，表 10 より，前後の文脈などの条件付きで適切としたものを含めれば，68.3％が対訳として利用可能なことが分かる．このことから，本手法の利用可能性が確認できた．

さらに，日本語での言い換えの成否が，ウイグル語対訳を得る場合にどのように影響しているかを調べた結果を表 11 に示す．この表から，日本語では言い換えに成功したもののうち，翻訳した結果が対訳語として不適切と判定されたものが，各品詞について，名詞 22.5％，動詞 22.4％，サ変名詞 20.8％の割合であったことが分かる．この原因としては，次の二点が挙げられる．一つは，日本語–ウイグル語機械翻訳システムが解析を誤ったために，正しい訳語が付与できなかったものである．例えば，「言い渡す → 命じる」は，日本語の言い換えとしては妥当であると判定した．しかし，翻訳システムが「命」を名詞として解析したために，正しく翻訳することができなかった．二つ目は，ウイグル語における単語の概念が異なるために，正しく翻訳できなかったものである．例えば，「出国する → 国を出る」は，日本語の言い換えとしては妥当である．しかし，「出国する」は，ウイグル語において「国 から 出る」と表現すべき単語であったために，対訳語としては不適切と判定された．こうした問題については，日本語–ウイグル機械翻訳システムの改善によって解決できると考えられる．

興味深い点としては，日本語の言い換えの評価としては失敗と判定されたにもかかわらず，そのウイグル語訳が，対訳として適切と判定されたものが少なからず存在したことである．その割合は，条件付き適切とされたものも含めた場合，名詞 46.7％，動詞 47.6％，サ変名詞 39.1％である．例えば，「打ち出す → 出す」という言い換えは，日本語の言い換えとしての評価では，「打ち」の部分の意味が欠落しているため，意味欠落と評価した．しかし，ウイグル語では「打ち出す」に相当するような単語がなく意味的には「出す」を翻訳した “ķoymaķ” が該当する．よって，この場合にはウイグル語としては適切な訳語が得られたことになる．

また，日本語における言い換え失敗の例として挙げた「書ける」に関しては，日本語では「書く」とは別の単語として辞書に登録されているが，使用した日本語–ウイグル機械翻訳システムでは「書ける」は「書く」に可能を表す接尾辞が接続した形であると解析し，「書ける」が

辞書になくても「書く」が辞書にあれば翻訳可能である．今回は，単純に辞書の未登録語をすべて対象としたが,「書ける」のように，未登録語であっても翻訳できる単語があり，こうした単語については今回の実験対象から除くべきであった．

こうした点を考慮すると，今回の言い換えにおける選抜段階での評価関数は日本語に注目しただけであったが，ウイグル語へ翻訳することを考慮した関数に変更することも考えらえる．

5　おわりに

本論文では，言い換え処理を収集段階と選抜段階の二段階に分け，収集段階に語義文ベースの手法を，選抜段階に類似度ベースの手法を適用することによる言い換え方法を提案した．さらに，獲得された言い換えを翻訳することによる日本語–ウイグル語対訳辞書の拡充も提案し，実験によりその利用可能性を確認した．

今後の課題としては，以下の項目が挙げられる．まず，日本語–ウイグル語対訳辞書拡充の効果に関する調査が必要である．本論文では，言い換えを用いて対訳辞書に新たな単語を登録したが，最終的な目的は，日本語－ウイグル語機械翻訳システムを用いて翻訳できる文数を増やすことである．よって，実際の文章が与えられたときに翻訳可能な文がどの程度増えるかといった評価や，文中の未登録語を動的に言い換えた場合の評価について調査する必要がある．

また，言い換え元に多義性がある場合の評価も必要である．本手法では，言い換え元を概念で区別しており，一つの語に複数の概念がある場合，それぞれについて別々の言い換えを獲得する．そうして得られた言い換えに対する評価と多義性解消への応用についても検討する．

さらに，本論文で扱った名詞，動詞，サ変名詞以外の品詞への適用も必要である．加えて，本論文の手法では名詞に関してはあまり結果が良くなかったため，現在，収集段階に類似度ベース，選抜段階に語義文ベースの手法を組み合わせた手法も試みている．

また，一度の言い換えでは翻訳できなかった単語については，翻訳できなかった部分を再度言い換えることによって解決できる可能性がある．例えば,「収納する」は「金銭を受納する」と言い換えられたが,「受納する」が翻訳できなかった．しかし，これをさらに言い換えて,「金銭を領収する」のようにすれば翻訳が可能になる．こうした多段階の言い換えについても検討し，その効果を確かめたい．

謝辞

本研究は，人工知能研究振興財団からの補助を受けて行われています．

参考文献

崔進，小松英二，安原宏 (1993). “EDR 電子化辞書を用いた単語類似度計算法.” 自然言語処理研究会 93-1, 情報処理学会.

藤田篤，乾健太郎，乾裕子 (2000). “名詞言い換えコーパスの作成環境.” 電子情報通信学会技術研究報告 TL 2000-32, 情報処理学会.

藤田篤，乾健太郎 (2001). “語釈文を利用した普通名詞の同概念語への言い換え.” 第 7 回年次大会発表論文集, pp. 331–334.

Hindle, D. (1990). “Noun Classification from Predicate-argument Structure.” In *Proceedings of the 28th Annual Meeting of the ACL*, pp. 268–275.

鍛治伸裕，河原大輔，黒橋禎夫，佐藤理史 (2002). “国語辞書とコーパスを用いた用言の言い換え規則の学習.” 第 8 回年次大会発表論文集, pp. 331–334.

釜谷聡史，小川泰弘，稲垣康善 (2002). “辞書語義文を利用した対訳辞書の拡充.” 情報処理学会第 64 回全国大会講演論文集 (分冊 2), pp. 91–92.

笠原要，松澤和光，石川勉 (1997). “国語辞書を利用した日常語の類別性判別.” 情報処理学会論文誌, **38** (7), pp. 1272–1283.

清瀬義三郎則府 (1989). 日本語文法新論–派生文法序説. 桜楓社.

黒橋禎夫，長尾真 (1998). 日本語構文解析システム KNP version2.0b6 使用説明書. 京都大学大学院情報学研究科, http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html.

黒橋禎夫，長尾真 (1999). 日本語形態素解析システム JUMAN version3.61 使用説明書. 京都大学大学院情報学研究科, http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html.

ムフタル・マフスット，小川泰弘，杉野花津江，稲垣康善 (2003). “日本語–ウイグル語辞書の半自動作成と評価.” 自然言語処理, **10** (4), pp. 83–108.

Murata, M. and Isahara, H. (2001). “Universal Model for Paraphrasing - Using Transformation Based on a Defined Criteria -.” In *Proceedings of 6th NLPRS Workshop -Automatic Paraphrasing: Theories and Apprications-*, pp. 44–54.

長尾真（編） (1996). 自然言語処理. 岩波書店.

日本電子化辞書研究所 (1996). EDR 電子化辞書仕様説明書. 日本電子化辞書研究所.

新村出（編） (1996). 広辞苑第四版 EPWING CD-ROM 版. 岩波書店.

小川泰弘，ムフタル・マフスット，外山勝彦，稲垣康善 (1999). “派生文法による日本語形態素解析.” 情報処理学会論文誌, **40** (3), pp. 1080–1090.

小川泰弘，ムフタル・マフスット，杉野花津江，外山勝彦，稲垣康善 (2000). “日本語–ウイグル語機械翻訳における派生文法に基づくウイグル語動詞句の生成.” 自然言語処理, **7** (3), pp. 57–77.

山本和英 (2001). “換言処理の現状と課題.” 第 7 回年次大会ワークショップ論文集, pp. 93–96.

略歴

小川 泰弘: 1995 年名古屋大学工学部情報工学科卒業. 2000 年同大学院工学研究科情報工学専攻博士課程後期課程修了. 同年より，名古屋大学助手. 博士

(工学)．自然言語処理に関する研究に従事．言語処理学会，情報処理学会各会員．

釜谷 聡史： 2001 年名古屋大学工学部電気電子情報工学科卒業．2003 年同大学院工学研究科計算理工学専攻博士課程前期課程修了．現，株式会社東芝．自然言語処理に関する研究に従事．情報処理学会会員．

ムフタル・マフスット： 1983 年新疆大学数系卒業．1996 年名古屋大学大学院工学研究科情報工学専攻博士課程満了．同年，三重大学助手．2001 年より，名古屋大学助手．博士 (工学)．自然言語処理に関する研究に従事．人工知能学会，情報処理学会各会員．

稲垣 康善： 1962 年名古屋大学工学部電子工学科卒業．1967 年同大学院博士課程修了．同大助教授，三重大学教授を経て，1981 年より名古屋大学工学部・大学院工学研究科教授．1997 年 4 月～2000 年 3 月工学研究科長・工学部長．2003 年 4 月より同大学名誉教授，愛知県立大学情報科学部教授．工学博士．この間，スイッチング回路理論，オートマトン・言語理論，計算論，ソフトウエア基礎論，並列処理論，代数的仕様記述法，人工知能基礎論，自然言語処理などの研究に従事．言語処理学会，情報処理学会 (フェロー)，電子情報通信学会 (フェロー)，人工知能学会，日本ソフトウエア科学会，IEEE，ACM，EATCS 各会員．

(2004 年 1 月 9 日 受付)
(2004 年 4 月 30 日 再受付)
(2004 年 7 月 5 日 採録)